УДК 004.93

Є.О. ГОФМАН, аспірант ЗНТУ, м. Запоріжжя,
А.О. ОЛІЙНИК, к.т.н., доц. ЗНТУ, м. Запоріжжя,
С.О. СУББОТІН, к.т.н., доц. ЗНТУ, м. Запоріжжя

СИНТЕЗ ДЕРЕВ РІШЕНЬ НА ОСНОВІ ВИБІРОК ДАНИХ З ПРОПУЩЕНИМИ ЗНАЧЕННЯМИ

Розглянуто завдання побудови моделей у вигляді дерев рішень на основі вибірок даних з пропущеними значеннями. Запропоновано новий метод ідентифікації дерев рішень, який використовує теорію функцій довіри для роботи в умовах невизначеності параметрів при розв'язанні завдання автоматичної класифікації за ознаками. Розроблений метод дозволяє ідентифікувати структуру та параметри дерев рішень. Табл.: 1. Бібліогр.: 9 назв.

Ключові слова: дерево рішень, ідентифікація, класифікація, невизначеність, функція довіри.

Постановка проблеми та аналіз літератури. При побудові моделей реальних об'єктів, процесів і систем часто виникають ситуації, коли в навчальній вибірці існують пропущені, суперечливі та аномальні дані. Причинами цього є: помилки при вимірі або введенні даних у навчальну вибірку, небажання респондента відповідати на деякі з поставлених питань, об'єднання не зовсім еквівалентних наборів даних [1, 2].

Більшість методів синтезу моделей на основі дерев рішень [2 – 4] у таких випадках не враховують екземпляри з пропущеними даними. Проте в такому випадку втрачається деяка інформація про досліджуваний об'єкт або процес, що може привезти до неприйнятної точності класифікації або прогнозування по синтезованому дереву рішень.

Тому актуальною є розробка методу синтезу дерев рішень на основі вибірки, отриманої в умовах невизначеності й неповноти даних.

Для розв'язання цього завдання в розробленому методі синтезу дерев рішень використовується теорія функції довіри, представлена в рамках передаваємої довірчої моделі (ПДМ) [5]. Така модель дозволяє створювати прийнятну систему класифікації завдяки здатності представлення неоднозначності. Крім того, ПДМ дозволяє експертам виражати часткові представлення більш гнучким способом, ніж це можна робити за допомогою функцій ймовірності. Такий підхід також дозволяє обробляти часткове або навіть повне незнання про параметри класифікації.

Мета статті – розробка методу синтезу дерев рішень на основі вибірок даних з пропущеними значеннями.

Основними завданнями роботи є:

– дослідження теорії функцій довіри;

– розробка методу ідентифікації дерев рішень за вибірками, що містять пропущені значення деяких ознак певних об’єктів;

– розробка програмного забезпечення, що реалізує запропонований метод.

Функції довіри. Нехай Θ – фрейм розпізнавання, що представляє кінцеву множину елементарних гіпотез, пов'язаних із проблемною областю. Множину усіх підмножин θ позначимо як 2^{θ}. Щоб представити ступені довіри, Shafer [6] ввів так звані основні розподіли довіри (basic belief assignments). Ці розподіли визначають частину довіри, яка покриває підмножину гіпотез, не покриваючи однозначної підмножини загальної множини при нестачі відповідної інформації [5]. Основний розподіл довіри (ОРД) – це функція m, яка приймає значення в межах [0, 1] для кожної підмножини A з Θ:

$$m : 2^{\theta} \longrightarrow [0,1], \tag{1}$$

при цьому:

$$m(\varnothing) = 0 \text{ та } \sum_{A \subseteq \Theta} m(A) = 1. \tag{2}$$

Підмножини A фрейму розпізнавання Θ, для яких $m(A)$ – однозначно позитивні, називаються фокальними елементами ОРД.

Ймовірність *Bel* та ймовірність *Pl* розраховуються за такими формулами:

$$Bel(A) = \sum_{B \subseteq A} m(B), \tag{3}$$

$$Pl(A) = \sum_{A \cap B \neq \varnothing} m(B). \tag{4}$$

Величина $Bel(A)$ виражає повну довіру, яка повністю передається підмножині A з Θ. $Pl(A)$ являє собою максимальна довіра, яка може покривати підмножина A.

У рамках теорії функції довіри легко виразити стан повного незнання. Це досягається за рахунок так званої порожньої довірчої функції, у якій єдиним фокальним елементом є сам фрейм розпізнавання Θ [6]:

$$m(\Theta) = 1 \text{ та } m(A) = 0, \tag{5}$$

за умови:

$$A \neq \Theta. \tag{6}$$

Одним з важливих понять теорії функції довіри – сполучення. Нехай Bel_1 та Bel_2 – дві функції довіри, що покривають дві різні частини знання. Нехай m_1 та m_2, визначають їх ОРД, відповідно.

Закон сполучення Демпстера спрямований на створення ОРД, яке представляє вплив об'єднаного доказу. Це визначене як [6]:

$$\forall A \subseteq \Theta, m(A) = (m_1 \oplus m_2)(A) = K \sum_{B,C \subseteq \theta : B \cap C = A} m_1(B) m_2(C), \quad (7)$$

де K – коефіцієнт нормалізації [6], $K^{-1} = 1 - \sum_{B \cap C = \varnothing} m_1(B) m_2$.

Правило сполучення Демпстера – коньюнктивне правило. Воно створює ОРД, якщо обидві частини правила є прийнятними. Подвійність цього коньюнктивного правила визначається диз'юнктивним правилом сполучення [7], яке створює ОРД, представляючи взаємодію двох частин доказу, коли відомо тільки те, що принаймні одна ОРД повинна бути прийнята, але не відомо, яка саме:

$$\forall A \subseteq \Theta, m_1 \vee m_2(A) = \sum_{B,C \subseteq \Theta : B \cup C = A} m_1(B).m_2(C). \quad (8)$$

Ці правила сполучення є комутативними й асоціативними. Таким чином, основний розподіл довіри, що визначається комбінацією декількох частин даних, може бути легко визначений шляхом багаторазового використання правила в будь-якому порядку їх застосування. Коньюнктивні та диз'юнктивні правила сполучення узагальнюють операції "АБО" й "ТА" теорії множин.

Проблема прийняття рішень у контексті ПДМ була вирішена в [6].

ПДМ заснована на двох рівнях інтелектуальних модулів:

– кредальный рівень, на якому переконання та висновки описуються функціями довіри;

– пігністичний рівень (pignistic level), на якому переконання використовуються для прийняття рішень і представлені функціями ймовірності, що називаються пігністичними ймовірностями.

Зв'язок між цими двома функціями досягається шляхом пігністичного перетворення, яке створює пігністичну функцію ймовірності *Betp*, що заснована на функції довіри:

$$BetP(B) = \sum_{A \subseteq \theta} m(A) \frac{|B \cap A|}{|A|} \text{ для всіх } B \subset \Theta. \quad (9)$$

Побудова дерев рішень з використанням теорії функцій довіри. Розроблювальний метод ідентифікації дерев рішень, заснований на

застосуванні математичного апарата теорії функцій довіри, характеризується особливостями як на етапі безпосередньої ідентифікації дерева рішень, так і на етапі класифікації екземплярів з використанням отриманого дерева.

Спочатку на етапі ідентифікації дерева рішень потрібно визначити основні параметри дерева у рамках теорії функції довіри, потім розробляється метод для побудови таких дерев рішень.

Запропонований метод для ідентифікації дерев рішень використовує теорію функцій довіри та заснований на розширеному алгоритмі C4.5 [3, 4], однак враховує при цьому невизначеність деяких параметрів, пов'язаних із завданням класифікації. Таким чином, виділяються деякі відмінності при визначенні гіпотез щодо цих параметрів при їхньому використанні.

Навчальна вибірка в рамках цієї невизначеної структури являє собою множину, що складається з елементів, представлених по парах (ознаки, клас), де для кожного екземпляра, як правило, відомі значення кожної з його ознак і унікальний клас, до якого належить даний екземпляр. Така навчальна вибірка (на відміну від тих, що традиційно використовуються для побудови моделей складних об'єктів і процесів) може містити дані, у яких є деяка невизначеність щодо класів. Іншими словами кожний клас навчальної вибірки може бути невизначеним або навіть невідомим, тоді як значення ознак, що характеризують кожний екземпляр, однозначно відомі.

Пропонується представити невизначеність класу будь-якого екземпляра шляхом основного розподілу довіри, заданому на множині класів. Це ОРД, традиційно задається експертом, являє собою переконання цього експерта про фактичне значення класу для кожного екземпляра в навчальній вибірці.

Серед переваг роботи з функціями довіри необхідно відзначити те, що легко можна виразити дві граничні ситуації (повне незнання та повне знання):

– якщо невідомою є будь-яка інформація про клас екземпляра, ОРД буде являти собою порожню функцію довіри:

$$m(\Theta) = 1 \text{ і } m(C) = 0 \text{ для } C \subset \Theta; \qquad (10)$$

– якщо клас екземпляра є однозначно визначеним, він буде представлений функцією довіри:

$$m(C_i) = 1 \text{ і } m(C) = 0 \text{ для всіх } C \neq C_i, C \subseteq \Theta, \qquad (11)$$

де C_i – одиничний клас.

Після того, як навчальна вибірка описана, слід задати другий важливий параметр розроблювального методу – міру вибору ознаки, яка буде використовуватися для вибору тестової ознаки в кожному вузлі дерева рішень.

Ця міра дозволяє визначити кількість сили диференціації кожної ознаки щодо кожного класу. За рахунок такого підходу досягається оптимізація дерева. Часто в якості такої міри використовується інформаційний критерій Квінлана [8, 9].

У рамках розроблювального методу міра вибору ознаки розширюється, і вона дозволяє працювати з невизначеністю, використовуючи теорію функції довіри. Нехай Bel_j – функція довіри, задана на множині можливих класів, що й описує переконання експертів про фактичне значення класу, до якого належить об'єкт I_j. Нехай S – підмножина екземплярів навчальної вибірки, з якої випадковим чином вибирається один екземпляр. Функція довіри, яка описує переконання про конкретний клас, до якого належить цей випадково відібраний екземпляр, є середньою функцією довіри, що приймає екземпляр в S. Тоді:

$$Bel_s(C) = \frac{\sum_{I_j in S} Bel_j(C)}{|S|}, \tag{12}$$

для всіх C підмножин з $\Theta = \{C_1, \ldots, C_n\}$.

Слід зазначити, що ОРД і пігністичні ймовірності, пов'язані із цією середньою функцією довіри, є пропорційними до основного розподілу довіри та пігністичних ймовірностей об'єктів з S (для будь-яких підмножин C з Θ):

$$m_s(C) = \frac{\sum_{I_j in S} m_j(C)}{|S|}, \tag{13}$$

$$BetP_s(C) = \frac{\sum_{I_j in S} BetP_j(C)}{|S|}. \tag{14}$$

Пропонується виконувати такі етапи для визначення інформаційної значимості ознак.

1. Розрахувати функцію середньої пігністичної імовірності $BetP_T$, виходячи з навчальної вибірки T. Потім розрахувати ентропію розподілу класів в T:

$$Info(T) = -\sum_{i=1}^{n} BetP_T(C_i)\log_2 BeP_T(C_i)\,. \qquad (15)$$

Ґрунтуючись на отриманих даних, розрахувати приріст інформації, забезпечений кожною ознакою A:

$$Gain(T, A) = Info(T) - Info_A(T). \qquad (16)$$

2. Для досягнення виконання попереднього етапу необхідно розрахувати $Info_A(T)$ для кожної ознаки. Пропонується застосовувати той же підхід, що використовується для розрахунків $Info(T)$, але обмежуючись множиною екземплярів, які характеризуються однаковим значенням атрибута A, і усереднюючи ці умовні інформаційні міри.

Таким чином, для кожного значення ознаки a_m виділяється підмножина T_m отримана з екземплярів T, для яких значення відповідної ознаки дорівнює a_m. Далі розраховується середня функція довіри *Bel*, після чого застосовується пігністичне перетворення для розрахунку пігністичної ймовірності *Betp*. Після отриманих перетворень можна розрахувати $Info(T_m)$, при цьому T_m являє собою навчальну вибірку, у якій значення ознаки A дорівнює a_m.

3. Шукане $Info_A(T)$ буде дорівнює зваженій сумі $Info(T_m)$ щодо розглянутої ознаки. При цьому пропонується, щоб $Info(T_m)$ були навантажені пропорційно значенням кожної ознаки в навчальній вибірці:

$$Info_A(T) = \sum_{m=1}^{k} \frac{|T_m|}{|T|} Info(T_m) = -\sum_{m=1}^{k} \frac{|T_m|}{|T|} \sum_{i=1}^{n} BetP_{Tm}(Ci).\log_2 BetP_{Tm}(Ci)\,. \qquad (17)$$

4. Після того, як обчислені інформаційні значимості ознак, вибирається ознака з найбільшою інформаційною значимістю.

Крім вибіркової міри ознаки, також повинні бути визначені два інші важливі параметри роботи методу:

– стратегія поділу: необхідно створити гілки для кожного значення ознаки;

– критерій зупинення: дозволяє припиняти розширення дерева та визначати вузол як лист. Таким чином, визначається, чи необхідно розділяти навчальну підмножину далі. У контексті розроблювального методу запропоновані такі варіанти критерію зупинення.

1. Якщо створений вузол покриває тільки один екземпляр, то даний вузол оголошується як лист, який характеризується тим ОРД, який уже визначений у навчальній вибірці.

2. Якщо вже немає подальшої ознаки для аналізу або якщо критерій інформаційної значимості для ознак, що залишилися, є меншим нуля, тоді вузол оголошується листом, де його ОРД буде результатом кон'юнктивного сполучення ОРД екземплярів, що відносяться до даного листа, розрахованого за законом Демпстера.

На відміну від традиційного дерева рішень, у якому кожний лист визначає унікальний клас, запропонований метод призначає кожному листу ОРД, виражаючи, таким чином, множину переконань про різні класи структури розпізнавання.

Нехай T – навчальна вибірка, що складається з екземплярів, що характеризуються ознаками $(A_1, A_2, \ldots, A_m)$, і які можуть належати множині класів $\Theta = \{C_1, C_2, \ldots, C_n\}$. Кожному об'єкту $I_j (j = 1, \ldots, p)$ з навчальної вибірки буде відповідати основний розподіл довіри, що виражає кількість переконань, які відносяться до підмножини класів.

Таким чином, розроблений метод містить етапи, описані нижче.

1. Генерація кореневого вузла дерева рішень, що включає всі об'єкти навчальної вибірки.

2. Перевірка задоволення поточного вузла критерію зупинення:

– якщо перевірка дала позитивний результат, то вузол оголошується листом і розраховується його ОРД так, як було описано вище;

– в іншому випадку – знайти ознаку з найбільшою інформаційною значимістю. Ця ознака буде розглядатися як корінь дерева рішень, пов'язаний з усією навчальною вибіркою.

3. Застосування стратегії поділу з метою створення гілки для кожного значення ознаки, обраної як корінь. Цей поділ веде до декількох навчальних підмножин.

Етапи 2 і 3 повторюються для кожної навчальної підмножини доти, поки вузол не буде оголошений як лист.

4. Зупинення побудови дерева рішень відбувається тоді, коли всі вузли останнього рівня дерева є листами.

Варто відзначити, що запропонований метод одержує такі ж результати як метод C4.5, якщо всі ОРД є однозначно визначеними. Така ситуація виникає, якщо клас кожного екземпляра з навчальної вибірки є унікальним і однозначно відомим.

Після того, як була зроблена ідентифікація дерева рішень, можна виконувати класифікацію екземплярів, що не відносяться до навчальної вибірки.

З однієї сторони запропонований метод може забезпечити традиційну класифікацію, при якій передбачається, що невизначені значення ознак будуть визначені. При такому підході, класифікацію слід виконувати в такий спосіб: починаючи з кореневого вузла й повторюючи перевірку ознаки на кожному вузлі, відбуваються відповідні переходи доти, поки не буде досягнутий лист. На відміну від традиційного дерева рішень, у якому кожному листу відповідає унікальний клас, у дереві рішень, побудованому за допомогою запропонованого методу, невизначені класи екземплярів визначаються за рахунок основного розподілу довіри, який відповідає досягнутому листу. Щоб одержати рішення і визначити ймовірність кожного окремого класу, пропонується застосувати пігністичне перетворення до основного розподілу довіри, і використовувати розподіл ймовірності, щоб розрахувати очікувану ефективність, необхідну для прийняття оптимального рішення.

З іншого боку, оскільки робота виконується з невизначеною вибіркою даних, розроблений метод класифікації дозволяє також класифікувати невизначені екземпляри, які характеризуються невизначеністю в значеннях їх ознак. У запропонованому методі передбачається, що нові екземпляри, які необхідно класифікувати, описані не тільки певними значеннями ознак, а також можуть характеризуватися невизначеними значеннями для деяких ознак. Крім того, навіть можуть бути ознаки з невідомими значеннями. Класифікація таких об'єктів полягає в пошуку листів, які можуть належати до розглянутого екземпляра, відслідковуючи всі можливі шляхи, викликані різними значеннями ознаки. У випадку невідомих значень, беруться до уваги всі розгалуження щодо розглянутої ознаки.

Як наслідок розглянутий екземпляр може належати до декількох листів, кожний з яких характеризується функцією основного розподілу довіри. Отримані ОРД повинні бути об'єднані, щоб одержати переконання щодо можливого класу екземпляра. Диз'юнктивне правило об'єднання, розроблене Сметсом [7], є прийнятним, оскільки воно припускає, що як мінімум один шлях є вірним. Так, у спрощеному випадку, в якому було виділено тільки два листи для розглянутого екземпляра, і клас екземплярів у першому листі A, а в другому – B, єдиним висновком із цього може бути те, що клас розглянутого екземпляра є або A, або B, тобто це і є диз'юнктивне правило.

ОРД, отримане на підставі диз'юнктивного правила, може бути перетворене у функцію ймовірності шляхом застосування пігністичного перетворення. Це дозволяє обчислити ймовірність приналежності розглянутого екземпляра окремому класу розглянутої проблемної області.

Експерименти та результати. Запропонований метод побудови дерев рішень на основі теорії функцій довіри був програмно реалізований у середовищі пакета Matlab 7.0.

Для експериментів використовувалася вибірка, що характеризувала випробування, проведені для визначення працездатності кузовів автомобілів [10]. Вибірка характеризувалася 46 ознаками, які описували стан 38 екземплярів. Ознаки характеризують значення зазорів і сполучень в 46 контрольних крапках, розташованих по всьому кузову автомобіля. При цьому в якості вихідних відгуків розглядалося 16 параметрів, що впливають на стан кузова автомобіля.

Таким чином, для кожного з вихідних відгуків були побудовані дерева рішень з використанням як запропонованого методу, так і з використанням методу CART. Після цього побудовані дерева рішень використовувалися для прогнозування значень вихідних відгуків на тестовій вибірці з розмірністю, аналогічною розмірності навчальної вибірки, яка характеризувалася невизначеністю значень деяких ознак. На основі отриманих значень вихідних параметрів для тестової вибірки були розраховані параметри роботи методів (усереднені значення) при прогнозуванні значень вихідних параметрів для тестової вибірки, які представлені в табл.

Таблиця

Результати роботи методів побудови дерев рішень

№	Метод ідентифікації дерева розв'язків	Усереднені характеристики роботи синтезованих дерев		
		Помилка прогнозування, %	Час роботи, с.	Кількість вузлів дерева, шт.
1	CART	4,3	57,1	35,1
2	Запропонований метод	1,2	52,3	32,3

Як видно з таблиці, дерева рішень, одержані з використанням запропонованого методу характеризуються меншою помилкою прогнозування та меншою складністю самого дерева, що сприяє кращій інтерпретабельності дерева.

Висновки. У роботі вирішено актуальне завдання автоматизації синтезу дерев рішень по прецедентах в умовах неповноти даних.

Наукова новизна роботи полягає в тому, що запропонований новий метод ідентифікації дерев рішень, у якому розраховуються пігністичні ймовірності віднесення екземплярів до класів на основі теорії функцій довіри, що дозволяє виконувати класифікацію екземплярів в умовах невизначеності або неповноти даних.

Розроблений метод відрізняється від існуючих методів ідентифікації дерев рішень фазою побудови дерева, оскільки враховується невизначеність, яка характеризує класи навчальних екземплярів за рахунок використання функцій довіри. Запропонований метод характеризується особливою процедурою класифікації нових екземплярів, значення ознак яких можуть бути невизначені, що дозволяє виконувати класифікацію неоднозначно заданих екземплярів.

Практична цінність отриманих результатів полягає в тому, що на основі запропонованого методу розроблено програмне забезпечення, яке дозволяє вирішувати завдання класифікації в умовах невизначеності або неповноти вихідних даних.

Список літератури: 1. *Субботін С.О.* Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник / *С.О. Субботін*. – Запоріжжя: ЗНТУ, 2008. – 341 с. **2.** *Лбов Г. С.* Анализ статистических данных с использованием деревьев решений [электронный ресурс] / *Г.С. Лбов, В.Б. Бериков*. – Режим доступа: http://math.nsc.ru/AP/datamine/decisiontree.htm. **3**. *Rokach L.* Data Mining with Decision Trees. Theory and Applications / *L. Rokach, O. Maimon*. – London : World Scientific Publishing Co, 2008. – 264 p. **4**. *Classification* and regression trees / *L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone*. – California : Wadsworth & Brooks, 1984. – 368 p. **5**. *Smets P.* The transferable Belief Model / *P. Smets, R. Kennes* // Artificial Intelligence. – 1994. – № 66. – P. 191–234. **6**. *Shafer G.* A mathematical theory of evidence / *G. Shafer*. – New Jersey : Princeton University Press, 1976. – 246 p. **7**. *Smets P.* Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem / *P. Smets* // International Journal of Approximate Reasoning. – 1993. – № 9. – P. 1–35. **8**. *Quinlan J. R.* Induction of decision trees / *J.R. Quinlan* // Machine Learning. – 1986. – № 1. – P. 81–106. **9**. *Quinlan J. R.* C.4.5: Programs for machine learning / *J.R. Quinlan*. – San Mateo : Morgan Kaufmann, 1993. – 312 p. **10**. *Гофман Е. А.* Использование деревьев решений для диагностирования автотранспортных средств / *Е.А. Гофман, А.А. Олейник, С.А. Субботин* // Информационные управляющие системы и компьютерный мониторинг: II Международная научно-техническая конференция ИУС и КМ-2011, 2011 г.: материалы конференции. – Донецк, 2011. – С. 159–163.

УДК 004.93

Синтез деревьев решений на основе выборок данных с пропущенными значениями / Гофман Е.А., Олейник А.А., Субботин С.А. // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2011. – № 36. – С. 35 – 45.

Рассмотрена задача построения моделей в виде деревьев решений на основе выборок данных с пропущенными значениями. Предложен новый метод идентификации деревьев решений, который использует теорию функций доверия для работы в условиях неопределенности параметров при решении задачи автоматической классификации по признакам. Разработанный метод позволяет идентифицировать структуру и параметры деревьев решений. Табл.: 1. Библиогр.: 9 назв.

Ключевые слова: дерево решений, идентификация, классификация, неопределенность, функция доверия.

UDC 004.93

Decision trees synthesis based on data samples with missing values / Gofman E.A., Oliinyk A.A., Subbotin S.A. // Herald of the National State University "KhPI". Subject issue: Information Science and Modeling. – Kharkov: NSU "KhPI". – 2011. – № 36. – С. 35 – 45.

The problem of constructing models in the form of decision trees based on data samples with missing values is considered. A new method of the identification of decision trees, which uses the theory of confidence functions to work in conditions of parameters uncertainty in solving the problem of automatic classification is proposed. The developed method allows us to identify the structure and parameters of decision trees. Tabl.:1. Refs.: 9 titles

Keywords: decision tree, identification, classification, uncertainty, belief function.

Поступила в редакцію 03.09.2011